DOCUMENT RESUME

ED 242 791 TM 840 213

AUTHOR	Burnham, Brian
TITLE	Use of Standardized Achievement Test Results in YRBE Elementary Schools.
INSTITUTION	York County Board of Education, Aurora (Ontario).
PUB DATE	Jul 83
NOTE	30p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Administrators; Basic Skills; Elementary Education; Elementary School Teachers; Foreign Countries; Grade 4; Grade 7; *Quantitative Tests; *Reading Tests; Research Methodology; School Counselors; Testing Programs; *Test Results; *Test Use; Use Studies
IDENTIFIERS	Canada; *Canadian Tests of Basic Skills; Test Batteries; *York Region Board of Education ON

ABSTRACT

Research was conducted to investigate the in-school use of results of the Canadian Tests of Basic Skills reading and mathematics batteries administered in the elementary schools of the York Region Board of Education (YRBE). Interviews were conducted with school administrators, guidance counsellors, and teachers. Test uses included reporting results to parents and staff and developing pupil profiles. Widespread uses of scores included identification of high achieving students, skill areas needing improvement, and "gratifying" results. Longitudinal comparisons, identification of needed learning resources, and comparisons with similar schools were less common uses of results. Several recommendations for board trustees and district administration to improve the testing program were made. Appendices outline characteristics of participating schools and the interview routine used in the study. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED242791

USE OF STANDARDIZED ACHIEVEMENT TEST RESULTS

IN YRBE SCHOOLS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. Burnham

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Brian Burnham
The Research Office
Division of Planning and Development
The York Region Board of Education

July 1983

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TABLE OF CONTENTS

Page

I. Summary 1

II. Why and how this study was conducted 2

III. The changing nature of the elementary school testing program

A. Overview 4
B. Historical Development 5

IV. Findings: overview 7

V. Findings: detail

A. System-wide uses 9

(i) Reporting to parents and pupils 9
(ii) Reporting to staff and subsequent action 10
(iii) Building a pupil profile 12

B. Widespread uses 13

(i) Identifying high achieving pupils 13
(ii) Identifying skill areas that need attention 13
(iii) Identifying "gratifying" results 15

C. Less common uses 15

(i) Longitudinal comparisons of results 15
(ii) Comparisons with similar schools 16
(iii) Identifying needed learning resources 16

D. Professional development 17

E. Different degrees of test usage 18

VI. Where do we go from here? Recommendations 21

Figure 1: Rated value of test results as a management resource 19

Appendix "A" --
Some characteristics of participating schools 25

Appendix "B" --
The Interview Routine 26

USE OF STANDARDIZED ACHIEVEMENT TEST RESULTS

IN YRBE ELEMENTARY SCHOOLS

I. Summary

The Canadian Tests of Basic Skills (CTBS) reading and mathematics batteries are administered annually to the region's grade 4 and 7 pupils. In May 1983, data on current in-school use of the CTBS results were collected from 15 YRBE elementary schools. Interviews were conducted in 13 of these schools with their principals plus a total of 24 other staff.

The relevant experience of these respondents is broad and deep. Some have known the board's standardized achievement testing program since it was implemented in 1972-73. Within the usual bounds of reliability for survey data, their testimony deserves credence. However, not all respondents were questioned about all matters under investigation. And certain points, issues, and questions were raised during the interviews that were not foreseen in the study design and hence were not fully dealt with. For such reasons, the following findings (however valid they are) may not be the last word on YRBE elementary schools' uses of achievement test results.

Collectively, CTBS test results are mainly used to help to:

(1) assess individual pupil progress and

(2) report it to parents;

(3) identify tentatively or to confirm the identification of pupils whose progress is "at risk" or pupils who should be in enrichment or gifted programs;

(4) place a new-to-school pupil in the most appropriate program or class;

(5) assess, by comparison with external norms, the collective performance of pupils as one means of assessing instructional effectiveness, either at the program level (e.g., mathematics) or the skills level (e.g., the components skills of mathematics problem-solving) or at both levels;

(6) identify priority program and professional development needs and

(7) identify learning material needed either for the classroom or school library;

(8) report to the community the children's collective success in achieving the basic skills of reading and mathematics.

Not all staff use test results for all these purposes. One very significant finding is that great variety in the use of the results can be found among schools. The main reasons for these differences seem to relate to:

(1) variations in perceptions of why region-wide group achievement testing was introduced, with concommitant variations in attitudes toward and value placed upon the testing program and in views of what schools were expected to do with results;

(2) the job held, e.g., teachers are mainly concerned with test results to the extent that they describe the achievement level of their individual pupils, and tend not to value results if there are no "surprises"; counsellors and vice-principals are more concerned with results which indicate needs for further testing, counselling or placement in another program, and tend to place more value on results which confirm "satisfactory" pupil progress; lead teachers are more likely to be interested in what the results say about program effectiveness for "their division"; principals have these concerns but are also interested in results across a grade and over the years and what the data say about instructional effectiveness and possible program or professional development needs, i.e., they have school management concerns as well as individual pupil or individual class management concerns;

(3) variations in respondents' experience, both in terms of the variety of roles they have held, in the sorts of schools and communities in which they have served, and the usefulness of test results to them in these various situations;

(4) the extent to which the respondents have correlated or synthesized test data with other indications of pupil progress so that each piece of information is used in the context of other data.

The full report has several indications of some further research which might be undertaken. Implications for trustees, administrators and school staff are also found throughout the report. A final section of this report, "Where do we go from here?" proposes a line of action based on a synthesis of indications and implications.

II. Why and how this study was conducted

This research attempts to answer the question, raised by trustees in January 1983, "How do schools use the results of the board's standardized

achievement testing program?" The context for this question was the perception that results were not being used advantageously in the schools and therefore thought should be given to discontinuing the program. (There may also be the implication that the results are not put to good use by trustees and senior staff.)

The original intent of the research was to answer the question for both the elementary and secondary panel. A design was developed and all principals were informed (February 1983) of pending in-school data collection. Subsequently, a random sample of 12 elementary schools was selected and visited in May.

Characteristics of the schools visited are given in Appendix "A".

However, difficulties with the administration of the testing program in the secondary panel delayed the reporting of their students' results until late in May. Since this was only the second annual administration of this particular test battery and since this year's results could scarcely be used normally, it was concluded that there would be little point in including the secondary panel in this year's investigation.* Secondary schools were informed that their schools would not be included in this year's study.

Elementary school interviews, conducted May 16 and 27 inclusive, were preceded by two communications. In an April 26 memorandum, all principals were reminded of the project's purpose and were invited to volunteer their schools if they were not among the sample chosen but wanted to make input. On May 5-6 principals of all schools chosen were informed that their staff were to be participants, given particulars of the in-school data collection procedures, and asked to inform staff of the details. Dates of school visits were arranged by mutual consent.

In addition to the 12 randomly selected schools, three others voluntarily filed written reports: one of these was visited and two hours of in-school interviews were conducted to flesh out their submission.

In each of the 12 "sample" schools, an interview was first conducted with the principal and some documentation collected and discussed. Principal interviews ranged up to nearly two hours, with a median average of about 50-60 minutes.

---

* Interruptions in high school programs in June are obviously unwelcome. The time which had been allotted for secondary school enquiries was used to interview senior administrators who were familiar with the program, with the results and their usage by schools or by the superintendency. Time was also used to interview trustees and school staff who volunteered to provide opinions and information. These supplementary data were integrated into the study report as applicable.

Two vice-principals, three guidance counsellors, three lead teachers, two teacher-librarians, and 14 "classroom teachers" (almost exclusively of grade 4 or 7 pupils) were interviewed for an average of about 25 minutes (some longer interviews ran up to 40 minutes). Since two principals, both vice-principals and all guidance counsellors and lead teachers also had regular classroom duties (often with grade 4 or 7 classes) and since two teachers each taught pupils from both grades 4 and 7, the respondents' experience bases were broad. Indeed, the within-school selection of interviewees was not random, but attempted to identify those with long and relevant experience. Nonetheless, one first-year teacher and one first-year principal were included in the poll.

The interview method used is summarized in Appendix "B".

In summary, even with a numerically small sample of elementary schools and teachers, there are good reasons to believe that one can take their observations to represent what generally would be heard if all schools were polled. Within the usual limits of reliability that attach to sample surveys, we have their picture of how the results of the board's mandated achievement testing program

- are being variously well used,
- why some variety in usage occurs, including minimal usage,
- what might be done if it is wished to foster fuller and more effective "group testing" programs.

It should be noted that while we have a reasonably reliable picture of school staff perceptions, their's is but one view of events. The perceptions of line supervisors and of the board's Testing Officer differ slightly on the nature and purpose of some events, and they mildly disputed some interpretations offered by those surveyed in this study when the present writer interviewed them during the course of this study or at the conclusion of the data gathering. In-school respondents differed among themselves, mainly on the particulars of what would constitute more valuable service (understandably, since they face different needs).

All this is to say that a rounded picture of the state of the region-wide testing program, including the value and use of results at the school level, is not easily arrived at. This report should not be the last word on all relevant matters. Rather, in many respects it should properly be viewed as raising as many questions as it presumes to answer, including as-yet-unasked questions about use of results at the system level.

III. The changing nature of elementary school testing programs

A. Overview

One cannot understand fully why there are different views on the usefulness of the YRBE's standardized achievement testing program unless one

first understands how and why the board-mandated program came into being and how it continues to impact on the school-level testing programs begun in the days of the predecessor boards. There are also factors peculiar to individual schools, their staff, or their community, that account for variations in estimates of the worth of standardized achievement testing. But most of these idiosyncratic features may likewise best be seen against this historic background.

A region-wide standardized testing program (commonly called "group testing") was introduced by the board in 1972-73. Initially the program attempted to serve several sets of information needs as expressed by some principals, guidance staff, administrators and trustees.

The purposes as well as administration of the program have changed over the decade. The principal motive of the changes seems to have been to provide senior staff and trustees with ever more timely, objective and pertinent information on "how well the system is performing". These changes have sometimes been seen by principals and others as undercutting group testing's more traditional (for school staff) instructional purposes related to individual pupils. Thus the current program, which involves only two elementary grades compared with four grades as in 1972-73, is less useful in the eyes of some school staff. In a document on standardized tests prepared for schools by the Testing Office, the section "Why Test?" gives five reasons. The first two relate to system purposes, the third to a school level goal, and the last two relate to individual pupils. It may be that this ordering of reasons is an important factor in the schools' perception of testing as essentially to serve system goals.

B. Historical development

However, these recent functional and material changes have, at the same time, put into the schools' hands valuable information about pupil performance not previously available. Schools are, in the main, unwilling to forego this information. They use it routinely, would deplore any curtailment of the program, but wish changes of various sorts to the region-wide program in order to serve better school, as distinct from system, needs.

Fifteen years ago, on the eve of the creation of the regional board of education, standardized ability and/or achievement and/or diagnostic testing was established in perhaps the majority of elementary schools. Before 1969, the major purposes of such testing related to class-level or school-level pedagogic decisions, e.g., pupil placement on the basis of strengths or weaknesses. A variety of tests was used, reflecting local decisions as to the relative merits of the tests which addressed school information needs.

In the early years of the new board, large scale testing programs, employing very new Canadian as well as U.S. tests, were instituted in the Primary Division in connection with research projects. In 1970-71 and 1971-72, about one-third of an entire age-grade cohort (the grade ones of

1970-71 who then became the grade two cohort of 1971-72) was given a new Canadian cognitive abilities test and also a new U.S. test of reading, spelling and mathematics achievement. Results were used in the two-year studies of pupil progess in the then-new modified "open plan/education" schools and also in the two-year study of the "early admission to kindergarten" scheme. The use of standardized ability, achievement, and learning style tests were also introduced into research studies in the grade 8 and 9 levels in 1972 and 1973.

The achievement results from these large pupil samples were thus available for examination by trustees and senior administrators. As those research studies were ending, the board, perceiving the value of a system-level performance achievement data source, decided that they wanted a regional program of achievement testing in the basic skills.

Beginning in 1972-73, a mandated elementary school reading and mathematics group testing program was put into effect with pupils (some exclusions) of grades 2, 4, 6 and 8, using a Canadian test battery. Later, there were changes in the elementary grades involved and secondary school students were included. Some testing options were introduced and then withdrawn. There were also changes in times when tests were administered and in the format and scope of the results fed back to schools. Thus there has been a decade of change, dictated centrally and not always well understood or wanted at the school level.

Currently, reading and mathematics achievement testing is carried out in grades 4, 7, 10 and 12 using the 1981 version of the Canadian test battery used since 1972-73.*

Workshops on the use of test results are offered for school staff. Some teachers and principals state that such programs have value, but complain that these programs have not addressed some of their other needs related to standardized testing. Notwithstanding this observation, the results are said to be useful in determining the comparative progress of the YRBE's elementary pupils at the end of the Primary and Junior Divisions and on other occasions. Useful comparisons, at an individual, class, school or system level, can be made using the national norms for individuals or for schools.

School staff differ on the relative merits of using a national or regional reference group (norm group) as the basis of comparisons. Most

---

* The Canadian Tests of Basic Skills (CTBS), Multilevel Edition, is used for elementary grades and the CTBS High School Edition is used with grades 10 and 12. The CTBS national norms describe typical performance by students from over 200 schools in 120 Canadian school districts. York Region and other Ontario boards were substantially represented in the standardization and special equating projects which produced the norms.

would like to have both sets of normative data, including school norms. Also wanted on the reports to schools are summative data on previous years' achievement levels. Many would like to have previous results repeated in order to facilitate year-over-year comparisons and estimates of improvement in program delivery. At present, previous years' records must be retrieved and tabulated by the schools before such comparisons can be made.

Before 1969, such comparisons were not possible, since different tests were administered at varying times and at different grades, all according to the different local perceptions of needs. No one recollects instances of "central" use of results pre-1969. There were fears, in 1972-73 and for some time thereafter, that the central administration would use results against the staff of schools whose pupils collectively performed at low levels. Principals (and supervisory officers who were consulted about this) could recall no use of these results in such an untoward fashion, however.

The new regional testing program disrupted some local testing practices and there was concern about displacement or elimination of existing school-level standardized testing. While this sometimes occurred, the present study turned up one school that has retained basically intact an "early identification" testing program of unknown antiquity (but is at least 12 years old) that has been integrated by the school with the board-wide program for a decade.

For schools and teachers with little or no experience with standardized testing, the changes which began in 1970-72 were sometimes viewed as threatening (just as the creation of a region board itself was often viewed as threatening to parochial patterns of operation). Maturation from these starting points has not always been rapid, but this survey found the "threat" to be largely a memory. However, attitudes to central motives in testing are now rooted in the view that the program needs to be altered (basically by additions) to make it more useful to pupils, their teachers and their parents.

IV. Findings: overview

The following uses of achievement test results were reported to be made by all or virtually all 15 elementary school principals from whom data were gathered:

1) reporting a pupil's achievement level to the pupil's parents;

2) reporting school achievement levels to all parents;

3) reporting school achievement in detail to staff, along with some analysis and, where appropriate, with requirement for further study or other action;

4) reviewing individual pupil results with relevant staff (e.g., "home room" teacher, guidance counsellor), mainly in the cases where current test results and evaluations of daily work at previous tests are significantly different, and determining whether further testing or other assessment procedure is needed;

5) updating pupils' Ontario School Records (OSRs) with the achievement test results. (Where results are viewed as unreliable, notes to this effect are added along with the questionable scores: in some few cases, the results are not entered for the same reason.)

The following uses of test results were reported to be made by principals or other staff in <u>many but not all</u> schools:

1) identifying high achieving pupils and determining whether they should be re-evaluated for enrichment or gifted pupil programs, if not already so placed;

2) identifying the test, sub-test, or skill area within a sub-test, where the school's results are below the national norm (or whatever score a school has taken as an indication of a deficiency which needs action) and analysing the result to determine whether individual pupils need help or whether the problem is widespread enough to focus attention on the programs and teachers involved as the areas needing development;

3) identifying "gratifying results which tell us we're on the right track" both for the morale boosting function and also (where applicable) to indicate "where changes in programs, methods, materials or staff are paying off".

The following uses of results were made in <u>a few schools only</u>:

1) previous years' results are compared, sub-test by sub-test or even skill area by skill area (within a sub-test) to determine change trends over several years and to decide what action is now appropriate;

2) results attained by similar schools are sought, comparisons made, and "reflected upon" (only the test and sub-test results are obtained and studied);

3) the teacher-librarian is involved in examining the implications of the results for materials aquisitions and for library programming (mainly for enrichment purposes as it is felt that collections were least adequate in this dimension).

In addition, there are other findings which cannot be tied to such quantifications: (1) professional development; (2) why there are different degrees and styles of test results usage.

V. Findings: detail

A. System-wide uses

Given that there are some differences in how schools approach shared objectives, we can nonetheless try to characterize their principal methods of using test results.

(i) Reporting to parents and pupils

All schools surveyed share the overall (aggregated) test results with parents or guardians of tested pupils as well as with the parent community at large.

The results attained at each grade level are sometimes announced at "Parents' Night" and usually published (e.g., in the schools' newsletters), but only in terms of the mean average performance of their pupils compared to the national norms. Percentile rather than grade equivalent data are reported. Regional norms (which are not provided) might be more appropriate basis for comparison. It was asserted by a few principals that using York Region as the reference group would be especially helpful in resisting the "I'm all right, Jack" syndrome. As a standard, they say, the national reference group may be inappropriate. The typically average York Region child or school has more educational advantages than the typical Canadian child or school and should therefore score higher, they believe.

The results attained by each child are privileged information. Individual's test performance is also less reliable than group performance. And parents are more likely to be more emotionally charged when dealing with their child's scores than with school average data. Therefore, principals and teachers take pains to ensure that a child's achievement scores (normally as percentiles: grade equivalent scores seem to be more readily misunderstood) are carefully explained to parents, preferably in an interview. Sometimes scores are sent out with the report card at term end. Normally, some written document is provided to explain what the scores mean (this document is often a local adaptation of materials forwarded to schools by the board's Testing Officer for this purpose).* Parents are invited to contact the teacher by phone for further clarification, if they want or need it and cannot attend the end-of-term interview session. One school asks parents to certify receipt of mailed-out results by signing and returning a tear sheet attached to the letter explaining how to interpret results.

---

* The kindness of Nelson Canada Limited, publisher of the CTBS, in permitting the YRBE to reproduce from their Teacher's Guide passages relating to test results usage, is hereby acknowledged.

Despite (mainly minor) variations in reporting procedures

-- no school reports experiencing anything in recent years by way of an "untoward incident" arising from sharing results ("It's now all routine", said one principal);

-- no school wants to discontinue receiving and sharing this sort of achievement information, though two principals state that the results only confirm what is already known by other means, especially with grade 7 students, and that testing in earlier years would be more useful (their teachers and staff at other schools share this view);

-- no school routinely reports test results directly to pupils (though their parents might), but discussion of these results and other relevant achievement data might be initiated by a teacher or counsellor with certain pupils.

It is not uncommon, when sharing information about pupil progress to arrange for a follow-up report to parents if the pupil is to be given further assessment (e.g., testing by the itinerant diagnostician).

The letters (or newsletters) which go to parents appear to the researcher to be both very sensitive to parents' concerns about their children's progress and very good at setting group testing into a proper perspective. Here, by way of example, is one cautionary paragraph from a total of nine by which one school tried to describe the testing program.

> "The testing program of your school is just one of many sources of information about your child's progress. While basic skills are important, there are many other aspects of development which are not measured by the tests: maturity, attitudes, interests, social relationships, esthetic understandings, etc. A close relationship with the school and a sharing of information is very helpful in contributing to your child's success and happiness in school."

(ii) Reporting to staff and subsequent action

In the usual course of events, principals study results when they arrive from the Testing Office. If the vice-principal is not involved in this initial perusual, he or she (where vice-principals are appointed) is then asked to study the results. Some discussion ensues as to further fact-finding, analysis, or sharing. However, if the principal finds "no surprises", sharing with all staff may follow immediately and informally. ("I left the printouts on the staff room tables and in the next days talked with the grade 4 and 7 teachers", said one principal: his judgment was that no one would find discrepancies between what the tests and previous teacher assessments said about the pupils.) Some principals present results to teachers formally, at the first staff meeting after results arrive in the school. This is common in schools with many staff.

Before sharing with staff, some principals and vice-principals engage in "red circling" certain data on the printouts of test results, especially low scores or distributions, for further attention. Previous test results for the school (one or both grades) may be drawn from files for comparison. A phone-around to determine by-grade by-test results of schools with similar characteristics is occasionally undertaken, e.g., if there is a dramatic swing in the scores for a battery or several sub-tests, to determine if similar circumstances occurred elsewhere. Invariably there is a scanning of individual scores, pupil-by-pupil, and identification ("red circling") of unexpected scores, or of "new-to-school" pupils' scores, or of scores that indicate "at risk" pupils, or of scores that indicate that the problem may be structural (i.e., with programs, materials, methods, or management) rather than just with individual youngsters. Marginal notes may be made then consolidated into a tentative plan for follow-up, perhaps first with the grade 4 and 7 teachers, or with all Primary or Junior Division staff.

While all this sounds time-consuming, principals affirm that because there are so few surprises this is not usually a big task. Teachers, whose work either follows upon the principal's or parallels it, say likewise. In most test administrations it is not usual to find more than one real "surprise" pupil per class. To find three is rare.

The greater surprise is to find that overall results in battery are not as high as expected. Often one sub-test (mathematics problem-solving was commonly cited) produces lower scores than acceptable for a substantial number in a class or grade, indicating shortfalls in instruction, at least relative to an external reference group. This will be dealt with later as it is not a situation found in virtually all schools.

Action subsequent to the identification, by test scores, of "at risk" pupils depends on many variables. Among these are (1) whether a trained counsellor is on staff -- and that is normally the case when there are grade 7 and 8 classes in the school, (2) whether the pupil is on a rotary timetable, and (3) whether the school enrolment is so small that staff communicate informally about pupils -- thus obviating formal case studies. Let us elaborate.

Where only a small number of pupils achieve an unexpectedly low score on a test or sub-test, the home room teacher (possibly assisted by the guidance counsellor when one is on staff) or the principal may convene a "case conference" involving staff who have taught any child identified by the CTBS as "at risk". Usually this is an informal exchange of information about the child and appears to occur only in large schools and only after it has been determined that the test scores are not the artifact of a non-academic origin (e.g., that the child was not simply ill during the testing, or upset due to a transient personal or family crisis).

The term "child" rather than "pupil" has been used to signify that this sort of procedure is far more common for the grade 4 pupil than for

the grade 7 youngster. By grade 7, as several interviewees put it, there are few surprises. Where these do occur, it is usually because of an unexpected and substantial drop in performance. A guidance counsellor (or vice-principal) will usually conduct an interview with the pupil to explore the situation. A consultation with the teacher(s) of the pupil may precede or follow that interview and a line of action jointly developed, if needed.

In one senior public school where grade 7 is the entry point, the CTBS results are important as they represent current and objective data to supplement the OSR's information on "new" pupils. If the test results suggest a pupil may be "at risk", a reporting sheet is circulated to each of the youngster's teachers to collect further and up-to-the-moment information on academic performance and behaviour. A case conference may then be in order, depending on the staff reports.

(iii) Building a pupil profile (OSR entries)

When results are sent to schools from the Testing Office they include synopsis of scores for each pupil, printed on pressure-sensitive adhesive strips that can be affixed in the appropriate section of the pupils' Ontario Student Records (OSRs).

All schools add these data to the OSRs, sometimes annotating the scores to indicate a belief that the scores (or more probably a few of the sub-test scores) may not be reliable indicators of performance capability. This is likely to occur for one or two sub-tests taken on a day when the pupil was upset or slightly ill, but tried the tests anyway.

Principals reported that in the past they were more likely to discard all the results if some were possibly lower due to such causes. Now, such action is relatively rare. There are two reasons given for this change. The first is a desire to preserve performance data to "build a pupil profile", as one respondent termed it, "even if that profile shows that test-taking upsets the kid". Secondly, there is awareness that test scores lower than expectation may indicate staff over-estimation of individual's achievement. Such may be the case of the younger, attractive, well-mannered, hard-working, articulate pupil who, nonetheless, has some unsuspected though minor learning difficulties. Teachers are influenced by "appearances" but standardized tests are not.

Again, there are few surprises with older pupils who have been in the school for several years. In general, it ought to be easier to detect a spurious score when dealing with older pupils. The guidance teachers interviewed on this matter state that follow-up with pupils and parents on suspected scores is common and often resolves doubts without further diagnostic testing. Where such testing is done, the entry of its results into the OSR allows future users of the "pupil profile" information to determine whether test scores are reliable for the pupil. (Note -- one respondent observed that the grade 7 pupil, typically age 12, is entering a life period where the body and mind are undergoing great changes, and

that stability in academic and social characteristics are not hallmarks of pupils in this period.)

B. Widespread uses

To this point we have described what is reported to occur in all, or virtually all, of the 15 schools from which data were collected. Other uses of these results are made by many (the majority, more or less) of the 13 schools where detailed information was gathered.

(i) Identifying high achieving pupils

As there is on-going identification of pupils who should have enrichment programs and as there is a standardized ability testing program (using the Otis-Lennon battery) to identify the academically gifted, the results of the CTBS testing at grades 4 and 7 yield few previously-undetected high achievers.

The exceptions include pupils transfering into the region, late bloomers", and youngster who have simply not been detected by the day-to-day sorts of assessments or the Otis-Lennon test. Again, there are said to be "few surprises" at grade 4 and even fewer at grade 7.

The usefulness of the CTBS in turning up pupils who should be added to enrichment programs or re-tested for placement in gifted pupil programs is not seen to be "great" by anyone. But apart from that, there is considerable difference as to its relative usefulness (1) in confirming previous assessments, (2) in more clearly establishing the relationship of achievement to ability -- the correlation is far from perfect, and (3) in spotting (or reconfirming) the pupil whose achievement is very high in mathematics but relatively low in reading (i.e., on ability tests, would score low average or low on the verbal but high or very high on the non-verbal scales) and who may need diagnostic testing to determine whether the pupil has a specific learning disability.

The differences in opinions about the usefulness of CTBS scores for casting light on these situations seem to reflect different patterns of looking for and at individual differences in pupils. This study was not structured to deal with this matter and some follow-up inquiries might be in order. It is the researcher's impression that most, but not necessarily all schools, lack expertise in the area of correlating various indications in pupils of ability/achievement that is well above but not uniformly distributed across, the basic skills.

(ii) Identifying skills areas that need attention

Sometimes results indicate that for one or more tests, sub-tests, or skills area within a sub-test, there is a general failure to achieve at whatever level the principal and staff deem appropriate. Problems then do not likely lie with the pupils, but with the program.

Many, but not all schools, report a routine for studying test results for evidence of structural deficiencies in their programs. That is, they look for indications that the course of study or the materials or the methods used may be the source of lower than acceptable results.

Schools which do not employ such a routine give various reasons: (1) the overall results of the battery and its various components meet or exceed their standard and they don't feel justified in tampering with success; (2) there has been great mobility in either staff or pupils and this renders such inquiries prone to failure, since it is impossible to get a fix on methods used; (3) the low results (e.g., in reading) trace to a simple origin (e.g., low vocabulary scores) and it is more profitable to put effort into vocabulary-building activities than to analyse specific reading skill deficiencies; (4) there are too few items per skill area to be definitive about what is or is not a real and general deficiency area, especially when the number who took the test is small (i.e., the reliability of the results is compromised by the joint effect of a small number of items and a small number of pupils); (5) by-skill-area results are not supplied by the Testing Office and have to be extracted by staff, which requires more time than it may be worth.

These conditions (whether or not they are valid) could change from year to year. For instance, greater staff stability could occur, or the Testing Office might be instructed to generate skill-area data.

Whatever the reasons for not using skill-area data, about half the schools do some analysis of school performance (which is often lower with respect to math concepts and problem-solving than staff find acceptable). These analyses may then lead to new plans for developing the deficient skills. In this respect, the materials from the CTBS Teacher's Guide (distributed by the Testing Office) provide a number of useful suggestions.

Respondents in these schools indicate: (1) that a plan is developed to change materials or methods -- especially the time and emphasis factors; (2) that more "break-out" of data by the Testing Office would help them to identify skill areas needing attention, just as item analysis by pupil, class and grade could be helpful in determining needed changes or special monitoring; (3) that it is sometimes hard to determine whether changes have been effective, since the CTBS is not given again to these pupils for three years (while they can and do develop their own criterion-referenced tests, teachers say they would like to administer the alternate form of CTBS to see, within the school year, how well remediation has worked: they can, of course, determine the effectiveness of changes in say, the Primary Division programs, through study of the test results in grade 4 in the following years).

While a few schools may have asked the Testing Office for a return of the item-analysis service, there is no indication that any have formally requested, through the Operations or Planning and Development Divisions, this service or for skills-area data. Perhaps this study should lead to a

formal polling of schools to determine which would use such information. (Note -- several principals have said that, in this age of computers, greater individualization of service should be economically feasible: in fact, individualization of test result reports existed in 1972-73 when analyses as well as scoring was almost totally by "manual" means and only aggregated results -- by school and for the region -- were computer processed.)

(iii) Identifying "gratifying results"

After program changes have been made, test result changes can be looked for in the achievement levels of subsequent cohorts of grade 4 or 7 pupils, as noted above. But when a school has gone the change route outlined in the previous section, it is very heartening to find success reflected in the results of the next year's (or years') testing program.

"It is one thing", said a teacher, "to see your own evaluations saying that now you've got it right, but quite another to see improvement confirmed by (CTBS) scores." Again, it is the objective and external validation that is a gratifying, morale-boosting reward for hard work.

As might be expected, the hard work does not always pay off, at least in the short run, for schools starting from rather low levels. Perhaps because some causes of low results, e.g., high pupil turnover, are beyond the school's control.

Some schools which have maintained fairly long histories of higher-than-average scores are very pleased with their results though they say they don't make much use of them. One principal felt that his staff were a bit too pleased and complacent about "their" scores and were somewhat unmotivated to try to imporve even further.

Several principals reported that sharing school level results with the parent community was a second opportunity to boost staff morale. Perhaps "looking good" by national norms leads to undeserved self-congratulation. As noted earlier, some feel that local norms rather than national should be used to set standards, as achievement at the 50 percentile national mark was not a real challenge for York Region schools, with all their advantages.

C. Less common uses

Looking now at the less frequent uses of standardized achievement test results, reported by at least one but not more than four schools, we find that these are "add-ons" to the more common applications.

(i) Longitudinal comparisons of results

Some principals compare current results with two or more previous years' records. By-test or by-sub-test results are even plotted to see if

there are trends. Conclusions are derived intuitively, not by formal trend-line forecasting formulae. Principals say that the comparisons are intended: (1) "To get the 'big picture', an overview of how the school program is working over the years (in the) basic skills"; (2) to get an "objective view of achievement in the Primary Division, where I've never taught and which I don't understand well enough to challenge my staff's view that the reading program is OK: the test results are some support of my feeling that the teaching of reading here isn't OK enough"; (3) to determine whether there are prevailing trends, up or down, in any sub-test, which might signal where changes are required or where continuing support is justified.

(ii) Comparisons with similar schools

The researcher expected to find that many schools systematically compared their results with those of schools serving similar or neighbouring communities. Only a few acknowledged taking more than a passing glance at those potential benchmarks. Rather, the national norm appears to be the standard of adequacy. And since most of the 13 YRBE schools visited compare satisfactorily or better, there was not much concern expressed about other results. (Note -- this matter was not pursued with the first schools visited nor some of the last schools.)

Principals and teachers generally took the view that they could not really determine whether other regional schools tested the same range of pupils, and hence comparisons might not be appropriate. When asked if they were sure that their school tested the same range of pupils as was included in the national norming and equating trials, they conceded that they could not. But, since they believe they test the fullest practicable range, omitting only those they are directed not to (e.g., English as a second language pupils), they feel that their own testees are not a screened population and would therefore resemble the typically average Canadian test group.

None could or would name a YRBE school that screened out pupils that ought to have been tested.

It might be useful, therefore, to clarify for elementary schools the intended test populations, just as circular memorandum 1982-83:97 did for secondary schools. To the extent that the use of local norms or other comparison data could be helpful in assessing school programs in relative and relevant ways, assurances of a common testing pattern is important. "Standardized" means that the identical test instrument is administered in the same manner to comparable pupil populations. The credibility of the process of using a reference group for comparison purposes rests on meeting all three conditions of the definition.

(iii) Identifying needed learning resources

In only two schools was it reported that the teacher-librarian was, by design, involved in studying test results to determine what centrally-

held learning resources might be called into fuller use, or what additional print or nonprint learning aids might need to be acquired to help address low achievement problems.

Some principals noted that they counted on the diagnostician to prescribe reading or other materials which should be tried with any child encountering substantial difficulties.

Most but not all principals who were asked did not believe that classroom teachers needed further help from teacher-librarians to identify learning aids. Nor did they see much need for printouts of selected remedial materials targeted to skills areas which might accompany the test results showing low achievement levels for some or many pupils.

Some principals said they thought something useful might result from either greater teacher-librarian involvement or from remedial materials compilations, but had not previously thought about these possibilities.

More thought, in fact, seems to have been given in recent years to ensuring that enrichment materials were available for pupils with high scores on either the ability or achievement tests.

D. Professional development

There is a miscellany of findings, concerns, and commentary that cannot be easily defined quantitatively, principally because few schools provided information (these matters were raised or evolved rather late in the interview cycle). These findings, etc., have been grouped for convenience under the two rubrics of professional development and differences in test results usage.

It appears that some professional development activities, leading both to broader understanding of test results usage and to trials of new approaches to test results use, would be appreciated by at least some schools. However, most principals interviewed on the broad question of professional development in testing said that they thought there were different needs for different schools. This study, they thought, might be a first step toward defining P.D. needs.

But they gave three warnings:

(1) while it is relatively easy to free principals or vice-principals or to attend workshops during the day (after school workshops of limited duration produce limited results, they say), it is increasingly difficult to release classroom teachers due to very limited supply teacher budgets;

(2) the principle of voluntarism has its limits and after 10 years of leaving it to staff to attend testing workshops on their own volition it is time to make certain training mandatory if the board is serious about effective use of test results;

(3) more of the sort of testing workshops that have been given in recent years may be required for new staff or for those who haven't yet wrapped their minds about the simplest uses of test scores: but more fundamental training is needed in test theory and construction and in results management and manipulation, perhaps using microcomputers and also the new techniques of exploratory data analysis, by "school leadership", including lead teachers and counsellors. (Note -- the "fundamentals" of test theory and construction were offered in the early 1970s but the all-day and half-day workshops were lightly attended.)

Some principals said they thought principals and vice-principals would prefer to be "trained" in testing apart from classroom teachers. Some thought that principals would be inhibited in a "mixed" group and would not pursue vigorously matters relating to use of test results for program evaluation.

One principal thought that principals and vice-principals, if properly trained, could direct in-school P.D. programs suited to their schools' particular needs.

E. Why there are different degrees of test results usage

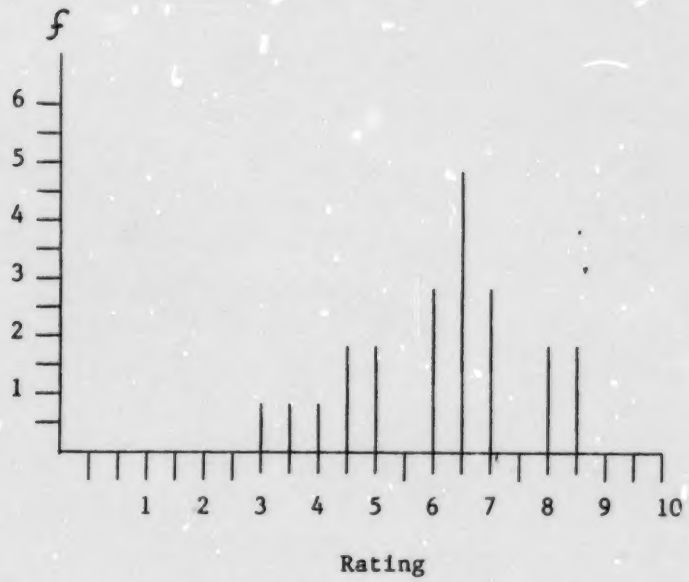
From an attempt by the principal of the third school visited to quantify his regard for the standardized test results emerged a question thereafter put to, but not necessarily answered by, almost all school staff interviewed: "On a scale of 0 to 10, the latter being the high score, how big a loss to your repertoire of school (or classroom) management resources would it be if the board ended its region-wide achievement testing program?"

In answering his own question, this principal gave an "8" rating. Only three other principals gave ratings this high (highest = 8 1/2). Most principals answered in the 6 to 7 range: some indicated that these scale positions would be higher if needed changes occurred. Only one principal (who teaches part-time in the school's grade 4 class) gave a below 5 score, a 4, in fact. (His grade 4 teaching partner gave a 3 score; the lowest rating given by anyone.)

Teachers gave ratings about 1 point (on average) lower than their principals, with only a few teachers valuing testing higher than their principals. This seems to reflect their narrower perception of the range of possible uses of test results: few teachers spoke of the use of results for program evalution, for instance.

The mean and median average scores fell at 6 1/2, but these ratings were not normally distributed. Instead, there were three "peaks" -- scores clustered about the 3 to 5 range, at 6 to 7, and the smallest peak at 8 to 8 1/2. (See Figure 1.)

Figure 1: Rated value of test results as a management resource



Again, these responses seem to indicate different current and position-oriented uses of test results, rather than judgments on the usefulness of test scores in other and all circumstances. That is to say, the question requires an answer based on the respondent's present situation. Since only 22 responded to the question, perhaps it is unwise to lean too heavily on this rating device.

To understand the differences in circumstances that are to be found across the schools surveyed (and hence across the region), some review of points made earlier needs to be worked into the following analysis.

Principals and teachers in schools serving fairly prosperous and stable communities report few surprises of any degree or with much frequency, especially in small schools. Where there is little teacher or pupil mobility, there is not much likelihood of a child moving from kindergarten to grade 4 (let alone to grade 7) without being a fairly well-known commodity. The few surprises are mainly to be found among the "quiet", "passive", "withdrawn", or that group who work long and hard to achieve what is more easily attained by their more gifted classmates.

Teachers report that they can spot most "bright" pupils within weeks of school opening because of their articulateness in class, their "track records" (including that of older siblings), and their performance on day-to-day learning tasks. How "bright" these pupils actually are is another question. Standardized ability and achievement tests help address that matter, especially since an external, objective yardstick is being used. A comparable situation exists for the slow learner. In both cases,

supplementary testing by the teacher, guidance counsellor, diagnostician, or psychologist, is considered much more definitive. But until the group testing turns up "surprises", the achievement capabilities of some pupils are unrecognized, perhaps because not all teachers can rise above their limited and subjective experiences of young people.

Generally, the less experienced classroom teachers valued test results more highly than any other respondent groups except guidance counsellors. Some principals commented upon the limits of subjective experience as associated not only with beginning teachers. One principal said that he did not appreciate the limits of his own rural school experience until he had his first posting to a suburban school. Nor did he fully appreciate the usefulness of standardized test results until he had to "get a fix" on hundreds of youngsters in a short time. The OSRs provided a lot of data, he said, and the school staff could help. But only the standardized test data in the OSRs and in the school's records served as understandable and objective measures of the achievement being attained in his new school. He now keeps yearly records of the school's achievement results and uses them with his staff as one means of quality control monitoring.

In situations where mobility (staff or pupils) is great, standardized test results are seen as more helpful. Principals of schools with higher than average turnover report that they (more-or-less routinely) study the OSRs of all transfer-in pupils, usually to assist in placement where this option exists. Although they may find teachers' comments of considerable use, the standardized test results often provide a more understandable standard than anecdotal or teacher-generated "marks" (especially letter grades) in reading and mathematics. Under these circumstances, the principal and other staff would like more standardized test results, they say.

One principal said that when he has surveyed the schools test results, he asks himself two questions "to get the juices flowing":

(1) What do the results suggest is needed by the kids and the staff?

(2) What further information do I need in order to confirm my initial needs assessment?

In answering his own question 2, he judges that the most useful additional information would include results of the CTBS Work-Study Test*

---

* Several principals asked whether this battery could be included in the regional testing program, at least as an option. As several stated the case, all staff are charged with helping pupils develop skills such as using reference works, getting data from maps, graphs, or tables, etc., but no one is responsible for assessing achievement of these learning skills in a comprehensive way. In fact, the Testing Officer has offered, as recently as two years ago, to add an administration of the CTBS study skills battery. There was not enough interest among principals at that time to justify instituting that activity.

battery, more information on pupil's attitudes to self and others, and (best of all) more test results from earlier years. Such information is just not presently produced for all children, though the early and continuing assessment program may in time generate a reasonable facsimile.

This principal feels that the value of any single datum on a pupil lies in the light it casts on all other data. Schools have to use all the pieces of information they can to understand and serve their pupils. Evaluating just one data source out of the context of this philosophy is likely to lead to underestimation (and underuse) of that data source. This principal feels that this survey may have done testing a disservice if it caused an evaluation "out of context".

This view was put to the few principals seen after hearing this position expressed. There was both agreement and elaboration. The gist of the matter is that, as a system, we still have some maturing to do before we learn to make good use of objective data in decision making. It was further suggested that this survey could be a stimulus to better test results use by trustees, senior administration, superintendents, curriculum committees and task forces.

VI. Where do we go from here? Recommendations

This study was undertaken to gather information about the in-school use of the results of the regional achievement testing program. This information was to be used by trustees in answering their question, "Should we continue the testing program?". The first alternative is to proceed now to an answer.

A second line of action is preferred by this study's author. Since the present program was instituted by the board (not the schools) in order to help trustees and administrators to determine the level of achievement of basic skills by pupils across the region, shouldn't trustees' and administrators' (rather than teachers') use of the results now be the prime concern?

If the trustees' question was reworded as "What needs to be done to improve the program?", there is still the assumption that the purposes of the program are understood by the decision-makers. School staff believe that the testing program was and is a board creation, certainly with some degree of usefulness for their own professional ends.

A possible line of action at this time would entail various "clarification" exercises, which follow as four recommendations.

It is recommended:

1. That trustees and senior administrators clarify the system uses of the test results, i.e., what they want from the program.

   This could include specifying what data they want (e.g., by-school by-skill reports as well as system-wide summaries) and how they want it formatted (e.g., three-year running averages for the system data, but by-school changes to be shown only as "up", "down", or "unchanged" year-over-year). They could also specify what uses could/should be made of the database by schools, support staff, supervisory officers and so on. Trustees and senior administrators might also decide whether national or local norms or both should be used as the reference (comparison) group and which (if either) should be the standard of adequate performance by schools, for internal use or publication.

2. That superintendents communicate to teaching staff the intentions that the system has for schools' use of test results and that principals, in turn, report school-based testing needs.

   These needs will vary but could include item analysis data and skills analysis data by-grade or by-pupil (or some variation, such as by-pupil only for low achieving youngsters). Some schools may also want other services to support "follow-up", such as directories of learning resources targetted to the skills areas covered by the tests. Schools may also want retrospective data in order to make longitudinal analyses. And some schools will probably want testing options added on to the system's program, e.g., the Work-Study Skills test battery, or the reading battery for grade 2, etc.

3. That some clarification of professional development programming needs be sought from the activities involved in 1 and 2, above.

   Further polling and consultation with school staff and supervisory officers may be needed.

4. That clarification of certain technical matters be sought.

   For instance, there should remain no doubt as to which elementary pupils must be tested and which may or may not be tested at a principal's option. Other technical matters to be addressed include improving scheduling of tests and distribution of materials and results (fluctuations create problems for schools). A clarification of what centrally-held testing materials are available to schools and better procedures for acquiring these are other technical matters that might be improved if desirable.

5. That the Testing Office be instructed to prepare costed alternatives for improving service to bring it into line with the new statements of the needs of the board and of the schools.

Some desirable changes may involve additional costs to the board. A superior level of service to schools could involve computerized test-scoring, either through the test publisher or by an in-house facility, and would require start-up or continuing expenditures. But the present and predecessor manual systems are labour intensive, hence neither fast nor cheap. Care should be taken, therefore, to look at costs (net costs in some cases) as well as benefits, before decisions to up-grade the system are taken.

Some desirable changes might actually reduce costs to the board. For instance, if a centrally-managed "optional, add-on testing program" (additional CTBS batteries or additional grades being tested) replaced the additional testing currently paid for out of school budgets, some economies might result from greater use of existing stocks of test booklets held by the Testing Office. The possiblity of saving money or, alternately, getting more from the dollars already being spent, should be examined as part of the process of rationalizing the current program.

These or similar actions would serve many ends if the wish is to improve the testing program. The responses from school staff interviewed during this study clearly indicate that virtually all want a better testing program, in the sense of better serving school needs. And the majority see the present program (and their schools' uses of results) as substantial enough, even with its limits, that its cancellation would be a considerable loss.

Achievement testing, or the broader field of pupil evaluation, has always been a major concern of educators but never so much as the last two decades. Evaluation is education's quality control function by which the degree of attainment of the enterprise's objectives is estimated. Many Ontario school boards are currently examining their achievement testing programs -- Peel has just completed a study and Halton began a study this May. The Ministry has funded extensive pools of test items that might help educators develop "criterion-referenced" tests to supplement standardized norm-referenced tests -- the Metro. Separate School Board is just one of many jurisdictions looking at computer-mediated use of such test item banks. The Ontario Institute for Studies in Education, continuing its many thrusts into this area, recently (1982) published What every teacher and parent should know about student evaluation, by Mark Holmes. The National Association of Secondary School Principals (USA) devoted much of its November 1982 NASSP Bulletin to standardized testing, with articles on results uses at various levels in a school system. The Canadian Education Association this year launched a long-term national study of student evaluation philosophies, practices, problems and prospects.

The motivation of such undertakings is to make improvements in the state of the art. It would seem premature, if not simply contrary-minded, for the YRBE now to walk away from defects in its present testing program.

It is recommended:

6. That a task force be established under the Superintendent of Planning and Development and charged with examining the future of standardized testing in the YRBE and of reporting to the Board not later than December 1984.

APPENDIX "A" -- Some Characteristics of Participating Schools

School	Area	Enrolment[a]	Staff[a]
I	1	<125	< 10
II	2	< 125	< 10
III	2	150-175	< 10
IV	3	176-200	10-15
V[b]	3	226-250	10-15
VI	4	251-275	16-20
VII[b]	3	301-325	16-20
VIII	3	326-350	16-20
IX	1	326-350	16-20
X	3	326-350	16-20
XI	2	351-375	16-20
XII	1	451-475	21-25
XIII	4	451-475	21-25
XIV	4	476-500	26-30
XV[c]	4	> 500	> 30

[a] To preserve the anonymity of these schools, pupil enrolment and staffing figures have been given as ranges. Note: "<" = less than; ">" = more than

[b] Volunteer school, written submission only

[c] Volunteer school, written submission plus interviews

APPENDIX "B" -- The Interview Routine

The research design called for each interview to begin with a single "focus" question. Follow-up "probe" questions were built upon the interviewee's initial response plus any data previously collected on the same subject. Three different focus questions were used in rotation in order to guard against casting each interview in the same mould. Use of the same opening question has a predictable tendency to produce similar and limiting lines of thought for both parties to the interview. This problem was obviated by this method.

This design recognized that any interview could not be exhaustive -- either in the sense of covering everything germane to the use of test results or in the sense of running on until one party or the other was physically or mentally exhausted. So only one focus question per interview was pursued, except where the respondent wished to address additional matters. Another practical consideration for some limiting of interview time was that teachers were being drawn out of their classes for these sessions, and principals and other staff from their various duties (many principals also found themselves covering the class of any teacher being interviewed).

The focus questions were:

(1) how are results used diagnostically and prescriptively to improve curriculum and instruction?

(2) how are results used with students and parents (e.g., as feedback and guide)?

(3) how are results used to determine professional development needs or what classroom and resource centre materials need to be obtained?

The list of follow-up probe questions grew as the interviews proceeded. These questions were often based on what had been reported by earlier respondents. Other information seeking or quantifying devices, such as rating scales (e.g., "On a scale of 0 to 10, the latter being the high score, how big a loss to your repertoire of management resources would it be if the board ended its region-wide achievement testing program?") were generated from points made by respondents.

Interviewees were informed that they could express views, ask questions, etc., after the interview. Many interviewees did not wait to the end to bring their own agenda to the fore.

Principals were asked to supply relevant documents, such as newsletters used to inform parents of the testing program. Such instruments were collected and may have later uses. At the time, they served to focus on specifics, to bring up relevant particulars, and sometimes to evoke

questions about the best means to announce or follow up test administrations.

The researcher encountered no instances of unwillingness to discuss any matter raised in the interview. Nor did he detect any lack of candour or other behaviours that would cast doubt upon the sincerity of the testimony. Indeed, for many respondents, the testing program was a routine they had known for as long as they had taught with the board. Most discussed the program matter-of-factly but were more animated when suggesting how the program could better serve school needs.

However, it needs to be remembered that the perception of the program is from one viewpoint, the schools', and that the program can also be assessed from the centre of the system, whose needs the program is presently and principally designed to serve. Nor was it possible to determine the factual basis of some statements. In the few instances where there was doubt in the mind of the interviewer about critical matters, other persons (including administrators and the current Testing Officer) were consulted for confirmation, if possible. It is clear that some differences of recall, etc., cannot be easily resolved. These few irreconcilable points are noted in the report of the survey.